НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

НАУЧНЫЙ ДОКЛАД

по результатам подготовленной
научно-квалификационной работы (диссертации)

# «Система обработки данных для космической обсерватории на основе мобильных телефонов для наблюдения за космическим излучение сверх-высоких энергий»

**ФИО: Борисяк Максим Александрович**

**Направление подготовки: 09.06.01 Информатика и вычислительная техника**

**Профиль (направленность) программы: 05.13.11 Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей**

**Аспирантская школа по компьютерным наукам**

Аспирант: _____ / Борисяк М.А. /

Научный руководитель: _____ / Устюжанин А.Е. /

Директор Аспирантской школы по компьютерным наукам: _____ /Объедков С.А. /

Москва, 2019

# Оглавление

# Глава 1

# Введение

Природа и источники космического излучения ультра-высоких энергий ($10^{18}$ эВ и выше) до сих пор остаются загадкой, как и механизмы стоящие за их ускорением ([1], [2], [3], [4]). Наблюдение таких частиц представляет особую ценность как для астрофизики, так и для физики в целом: так как ни одна теория еще не была сопоставлена с наблюдениями для энергий такого порядка, получение данных о частицах ультра-высоких энергий представляет большой интерес для научного сообщества: данные такого рода затрагивают даже экзотические концепции как идею дискретной структуры вселенной [5][1].

Сами частицы крайне затруднительно напрямую наблюдать на практике. Но, проходя через атмосферу Земли и соударяясь с молекулами воздуха, такие частицы порождаю вторичные частицы, которые в свою очередь могут породить следующее поколение частиц, таким образом создавая «ливень» вторичных частиц. Основная техника наблюдения за космическим излучением ультра-высоких энергий состоит в детектировании «ливня» вторичных частиц, который в зависимости от исходной энергии частицы и чувствительности сенсоров наблюдаем в радиусе около 0.25-1 км [4]. Имея достаточное количество данных о «ливне», можно оценить параметры исходной частицы, в том числе, энергию и направление.

Статистика наблюдений, содержащая место, время, энергию и направление космического излучения ультра-высоких энергий представляет собой важные данные, например, позволяющая указать на возможных кандидатов на источники этого излучения.

Основную трудность для исследований создает редкость появления таких частиц (начиная от одной частицы в год на квадратный километр), что требует либо большого времени наблюдения, либо большой площади покрытия сенсоров, что требует значительных денежных затрат.

Совсем недавно было замечено, что камеры мобильных телефонов способны выступать в роли детекторов элементарных частиц, в том числе вторичных частиц «ливня» ([6, 7]). В связи с этим был предложен альтернативный более дешевый подход к наблюдению за частицами ультра-высоких энергий — использование камер мобильных телефонов добровольцев из различных регионов планеты как части огромной распределенной обсерватории (эксперимент Cosmic RAYs Found In Smartphones, CRAYFIS [8]). Основная идея состоит в том, что в случае частиц ультра-высоких энергий неточный, но покрывающий огромную площадь сенсор становится более эффективен нежели высокочувствительные обсерватории с относительно низкой площадью покрытия. Поэтому, имея определенное количество активных телефонов в области «ливня», такой распределенный сенсор имеет гораздо больший шанс детектирования частиц ультра-высоких энергий, конечно, теряя при этом в точности оценок на параметры исходной частицы. Другим важным преимуществом является стоимость проекта, которая несравнимо ниже стоимости постройки обсерватории с тем же шансом детектирования. Для достижения чувстви-

[1] Здесь стоит обратить внимание на ограничения выраженные в единицах энергии

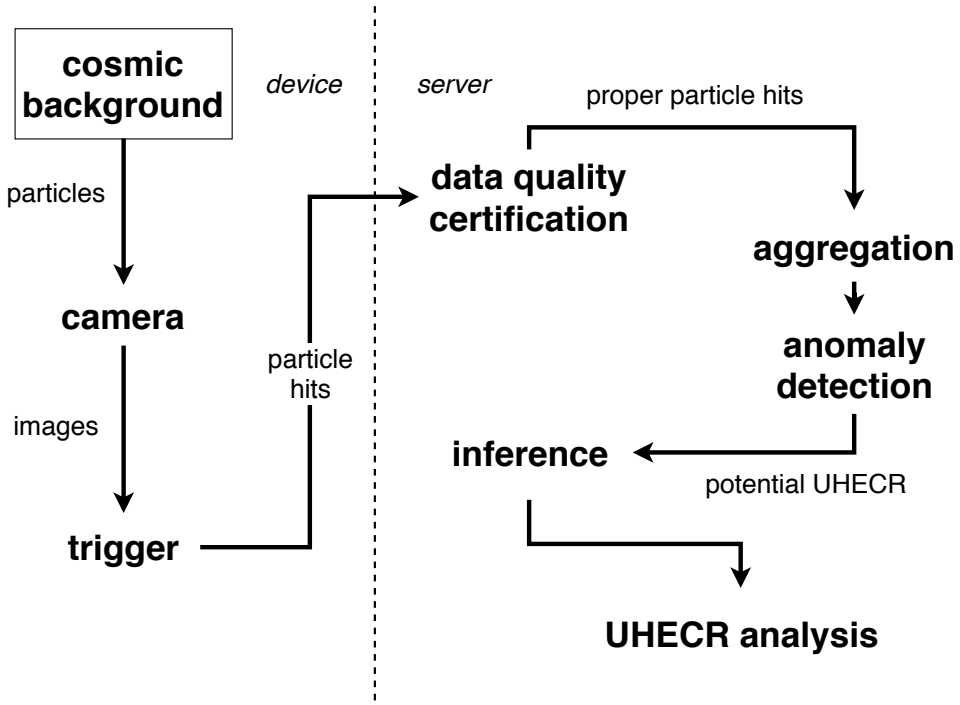



Рис. 1.1: Схема системы обработки данных эксперимента CRAYFIS. Cosmic Background — космический фон; camera — камеры мобильных телефонов, участвующих в эксперименте; trigger — нахождение следов частиц на изображениях, полученных камерами телефонов; data quality certification — процесс оценки качества приходящих данных; aggregation — агрегация данных с различных устройств; anomaly detection — детектирование отклонений в локальном поведении групп телефонов; inference — оценка параметров камер телефонов при помощи симуляции, определение типов и энергий частиц, участвующих в событии; UHECR analysis — определение наличия «ливня», оценка параметров космического излучения ультра-высоких энергий.

тельности, необходимой для того, чтобы различить космический фон от «ливней» космического излучения ультра-высоких энергий, количество волонтеров должно достигать миллионов [8].

Подробное описание проекта с точки зрения астрофизики и оценки на необходимую плотность активных телефонов можно также найти в [8].

На рисунке 1.1 приведена схема системы обработки данных. Эксперимент просит волонтеров, участвующие в эксперименте, включить специальное приложение эксперимента и оставить мобильный телефон камерой вниз, таким образом, чтобы свет на проникал в камеру. Космическое излучение, проходя через камеру телефона, способно активировать пиксели камеры на своем пути, создавая т.н. трэк (англ. track), позволяя таким образом определить факт прохождения частицы, и, в некоторых случаях, предсказать тип частицы, исходную энергию и направление.

Уже на этом этапе можно выделить три основные особенности в обработке данных. Первая особенность заключается в том, что различные камеры телефонов имеют различные и априори неизвестные характеристики, основной из который является глубина пикселей. Эти характеристики влияют на отклик камеры при прохождении частиц через нее (несложно заметить, что чем больше глубина матрицы камеры, тем, в среднем, выше длина следов частиц и ярче отклик).

Вторая особенность заключается в объеме данных получаемых с камеры телефона. Для достижения необходимого временного разрешения, частота съемки должна быть около 10 Гц, при этом частота наблюдения частиц космического фона через камеру составляет около одной частицы в минуту. Дополнительно, стоит заметить, что обычно частицы оставляют следы все-

го в несколько пикселей, в то время как современные камеры содержат миллионы пикселей. Для большей части космического фона, детектирование частиц решается простым порогом на активацию пикселей камеры, однако, космический фон также содержит минимально ионизирующие частицы (в основном мюоны), которые также составляют значительную часть «ливней», порожденных космическим излучением ультра-высоких энергий. Минимально ионозирующие частицы оставляют слабый отклик на камере телефона, который, в большинстве случаев, можно отличить от шума камеры только по форме следа частицы. С учетом низкой вычислительной мощности телефона, для детектирования подобных частиц требуется быстрый алгоритм нахождения трэков (триггер).

Третья особенность заключается в обучении триггера — получение разметки для реальных данных затруднен, так как не существует достоверного надежного способа подтвердить факт прохождения частицы через камеру телефона. Похожая задача часто возникает в физике высоких энергий, где для ее решения используют контрольные наблюдаемые.

После обработки триггером, потенциальные трэки частиц отправляются на сервера эксперимента, где происходит следующий этап обработки данных — определение качества данных. Так как у эксперимента нет контроля за мобильными телефонами, для дальнейшего использования данных, необходимо удостовериться в соблюдении условий проведения эксперимента. Например, камера устройство может быть не полностью закрыта и небольшое количество света проникает на матрицу. В этом случае, будет наблюдаться повышенная по сравнению с уровнем космического фона частота срабатывания триггера. К сожалению, формальные критерии качества данных получить затруднительно (в основном из-за разнообразия характеристик камер), и оценка качества данных предполагается осуществлять вручную, опираясь на статистики мотивированные предметной областью. Стоит заметить, что первый этап оценки качества данных должен происходить индивидуально для каждого телефона, что требует громадного количества работы (напомним, что для успешного эксперимента требуются миллионы волонтеров). Похожая проблема возникает при оценки качества данных экспериментами Большого Адронного Коллайдера. В данной работе рассматривается ряд алгоритмов нацеленных на автоматизацию оценки качества данных.

После проверки качества данных, поступающих с индивидуальных телефонов, данные агрегируются по времени наблюдения и координатам. Похожая на предыдущую задача возникает на этом этапе — нахождение аномальных агрегированных событий, но в отличии от верификации качества данных, основная цель этого этапа нахождение событий, кандидатов на «ливни». Для обнаружение аномалий предполагается использование аналогичного алгоритма, как и в предыдущем случае, однако, в данном случае (из-за редкости «аномальных» событий) требуются методы обнаружения аномалий, способные, с одной стороны, принять во внимание известные аномалии (предположительно полученные из симуляции), с другой, способные детектировать аномалии, непредвиденные заранее. С этой целью был разработан специальный метод для определения аномалий.

После отбора кандидатов на «ливни», порожденные космическим излучением сверх-высоких энергий, требуется детальный анализ события, с целью подтверждения искомого события и оценки свойств изначальной частицы. Для осуществления данного шага требуется точное определение свойств камер телефонов участвовавших в событии. Это возможно достичь, используя историю событий наблюдаемых устройством и точную симуляцию, путем нахождения таких параметров симуляции, которые восстанавливают распределение событий наблюдаемых устройством. Для нахождения параметров симуляции, предлагается использовать методы на основе состязательных сетей (Adversarial Optimization), адаптированные на случай не дифференцируемого симулятора. Основная проблема данных методов состоит в больших количестве вычислительных ресурсов, требуемых для симуляции необходимого количества событий. С це-

лью ускорения оптимизации, был предложен новый метод.

## 1.1 План диссертации

В начале работы кратко приведена необходимая информация о физических свойствах космического излучения ультра-высоких энергий, а также следствия, влияющие на анализ данных.

Далее, описаны алгоритмы калибровки камер мобильного телефона как детектора космического излучения, а также разработанный метод для получения эффективных алгоритмов калибровки как задачи поиска выбросов в статистических данных.

Следующая глава посвящена анализу изображений, получаемых с камер мобильных телефонов. Описаны алгоритмы обнаружения следов космического излучения на изображениях, а также специальный класс алгоритмов машинного обучения, основанный на Artificial Neural Networks, для детектирования редких событий в условиях ограниченных вычислительных ресурсов (в данном случае, вычислительных ресурсов мобильного телефона).

Следующая глава посвящена восстановлению параметров (типа частицы, энергии и углов падения) наблюдаемых с помощью камер мобильных телефонов частиц по отклику камеры. Основной задаче в этой главе является работа с данными физических симуляций, рассматриваются методы удаления систематической отклонений данных физических симуляций от реальных данных методами машинного обучения.

В последующий главах, рассматривается задачи детектирования и восстановления параметров космического излучения ультра-высоких энергий по событиям от индивидуальных мобильных телефонов. Основное внимание уделяется работе с потоками больших данных и распределенным вычислениям. Также рассматриваются задачи предварительной фильтрации потока данных, обнаружения редких событий, построения эффективных алгоритмов с помощью методов машинного обучения, аппроксимирующих точные статистические методы.

# Глава 2

# Обзор исследований по данной проблематике

Задачи в эксперименте CRAYFIS можно условно разделить на три класса: относящиеся к астрофизике и физике элементарных частиц, относящиеся к областям статистики и машинного обучения.

К первому классу задач относятся:

- постановка физических экспериментов;
- построение симуляций физических процессов;
- обоснование моделей и методов;
- получение экспериментальных данных.

Однако, эти задачи выходят за рамки данной работы. Для ознакомления с физическими аспектами см. [8] и соответствующую литературу.

## 2.1 Статистика и машинное обучение в астрофизике и физике высоких энергий

Эксперимент рассматриваемый в данной работе является астрофизическим экспериментом, однако задачи возникающие в нем часто являются характерными и для физики высоких энергий, поэтому с точки зрения анализа данных эксперимент можно рассматривать либо как астрофизический, либо как эксперимент из области физики высоких энергий в зависимости от решаемой задачи.

Статистические методы являются базовым инструментом анализа данных в физике высоких энергий. Так как физика высоких энергий является фундаментальной наукой, основным отличием от статистики в других областях является предпочтение отдаваемое методам частотного анализа, в отличии от байесовских методов, часто используемых, например, в астрофизике. В основном, используемые в физике высоких энергий и астрофизике статистические методы метод являются общепринятыми, за исключением, возможно, некоторых специфических для этих областей задач (эти задачи возникают также при применении машинного обучения и будут рассмотрены более подробнее далее). Описание основных методов в физике высоких энергий можно найти, например, в [9], основы баейсовского анализа изложены например в [10].

Однако, в данной работе в основном используются подходы машинного обучения, которые набирают популярность, в том числе, в прикладных и фундаментальных науках (например, в астрономии [11]), машинное обучение активно применяется в экспериментах CERN LHC.

Одним из примеров использования машинного обучения является система фильтрации событий в эксперименте LHCb CERN. Как и многие современные физические эксперименты (в том числе и CRAYFIS), в LHCb доля интересующих событий в потоке данных крайне мала, а поток всех событий слишком велик для полной обработки и даже для хранения. Поэтому многие эксперименты используют системы фильтров событий (триггеров), отсекающие неинтересные с точки зрения эксперимента события [12], [13], [14]. При этом один из триггеров построен на основе Bonsai Boosted Decision Tree [15], [16], одного из популярных алгоритмов машинного обучения.

Широко известно соревнование по построению алгоритмов машинного обучения для обнаружения бозона Хиггса [17], проводимое другим LHC экспериментом ATLAS. Практически все решения строились на основе методов машинного обучения, таких как Boosted Decision Trees и Artificial Neural Networks [18].

В эксперименте CRAYFIS исходными данными являются изображения с камер телефонов. Одними из основных методов машинного обучения для изображений является Convolutional Neural Networks (см., например, [19]), одна из разновидностей Artificial Neural Networks.

Активно развивающаяся область машинного обучения Deep Learning (обычно используются Deep Neural Networks [20]) также используется в физике высоких энергий и астрофизике. В качестве примера, можно привести сложную задачу поиска «экзотических» частиц [21], в которой авторы применяют известные техники Deep Learning для тренировки Deep Artificial Neural Network. Полученный классификатор превзошел популярные алгоритмы машинного обучение, в том числе и Boosted Decision Trees.

В фундаментальных науках зачастую невозможно использовать реальные данные для обучения так как настоящие метки для реальных событий неизвестны. Для обучения часто используют данные, полученные по методу Монте-Карло специальными симуляциями, которые воплощают современное понимание фундаментальных законов. Этот момент вносит дополнительные трудности в применении машинного обучения. Например, чтобы измерить неизвестную массу бозона Хиггса, требуется фильтр инвариантный к его массе. В работе [14] рассматривается задача построения таких фильтров для физики высоких энергий. Другим следствием такого подхода часто является разница в распределениях параметров между реальными и симулированными данными. Для решения этих проблем применяют техники перевзвешивания (статистический подход к перевешиванию можно найти в [9]), которые ведут к созданию новых алгоритмов и методов: [22], [23].

Обзор применяемых в астрономии методов машинного обучения и возникающих задач можно найти в [11].

Также краткий обзор соответствующей литературы приведен в каждой главе.

# Глава 3

# Обнаружении и фильтрация событий на мобильном телефоне

Одной из особенностей эксперимента CRAYFIS является колоссальный поток данных, получаемый всеми сенсорами: для грубой оценки потока мы можем воспользоваться следующими рассуждениями: 1 миллион мобильных телефонов, каждый из которых делает мегапиксельный снимок 30 раз в секунду, производит 30 Тбайт сырых данных в секунду. Однако, понятно, что лишь малая часть этих данных содержит какую-либо ценную информацию. Эксперименты и симуляции GEANT показывают, что космическая радиация редко оставляют следы более 10 пикселей и ожидаемый поток составляет около 1 частицы в минуту. Обрабатывать всеь поток сырых данных не представляется возможным, как минимум, из-за ограничений на передачу данных. Однако современные мобильные телефоны обладают относительно высокой вычислительной мощностью. В связи с этим первичная фильтрация событий осуществляется на устройстве пользователей.

Это, однако, все еще является проблематичным — вычислительной мощности мобильных телефонов не хватает для запуска стандартных алгоритмов машинного зрения, таких как сверточный нейронные сети. Как уже было замечено выше, ожидается, что космическая радиация активирует только малый участок камеры раз в минуту. С другой стороны из-за очень низкой ожидаемой плотности участников, от алгоритма фильтрации требуется очень высокая точность для достижения адекватной чувствительности обсерватории. В связи с этим встает необходимость в алгоритме, который с одной стороны достигает качества стандартных алгоритмов машинного зрения, с другой способен быстро отбрасывать шумовые участки снимков. Для алгоритмов такого рода часто используют термин ”триггер”.

Одним из подобных алгоритмов является каскады Виолы-Джонса, описанные в следующей секции. Среди альтернативных подходов стоит выделить R-CNN, Fast R-CNN [24] и Faster R-CNN [25]. Стоит заметить, что перечисленные требуют запуска отдельной сети для определения регионов-кандидатов, что, с учетом низкой производительности мобильных устройств, уже является недостижимым для эксперимента CRAYFIS.

## 3.1 Каскады Виолы-Джонса

Каскады Виолы-Джонса являются одним из первых широко применяемых на практике общих алгоритмов машинного зрения. Каскады Виолы-Джонса строятся на основе фильтров Хаара, который задается прямоугольником с координатами $(x_0, y_0, x_1, y_1)$:

$$\text{conv}(I, x_0, y_0, x_1, y_1) = \sum_{x_0 \leqslant x \leqslant x_1} \sum_{y_0 \leqslant y \leqslant y_1} I[x, y]; \tag{3.1}$$

где $I$ — обозначает двумерный массив, задающий изображение.

Несмотря на простоту по сравнению с классическими фильтрами, фильтры Хаара обладают одной уникальной особенностью — после предобработки, сложной вычисления (3.1) составляет $O(1)$:

$$\text{conv}'(J, x_0, y_0, x_1, y_1) = J[x_1, y_1] + J[x_0, y_0] - J[x_1, y_0] - J[x_0, y_1]; \tag{3.2}$$

где:

$$J[x, y] = \sum_{1 \leqslant i \leqslant x} \sum_{1 \leqslant j \leqslant y} I[i, j]. \tag{3.3}$$

Легко убедится в том, что массив значений $J$ может быть получен за линейне время по следующим формулам:

$$J[1, 1] = I[1, 1]; \tag{3.4}$$
$$J[x + 1, y] = J[x, y] + I[x + 1, y]; \tag{3.5}$$
$$J[x, y + 1] = J[x, y] + I[x, y + 1]. \tag{3.6}$$

Каскады Виолы-Джонса используют этот факт, для построения быстрого алгоритма распознавания изображений. Каскад состоит из последовательности фильтров и порогов: $\{(x_0^i, y_0^i, x_1^i, y_1^i), \tau^i\}_{i=1}^N$ и работает по следующему принципу:

$$f(I) = f^N(I); \tag{3.7}$$
$$f^1(I) = 1; \tag{3.8}$$
$$f^i(I) = \begin{cases} f^{i-1}(I), & \text{если conv}(I, x_0^i, y_0^i, x_1^i, y_1^i) > \tau^i; \\ 0, & \text{иначе;} \end{cases} \tag{3.9}$$

где 0 обозначение отсутствие искомого шаблона в изображение, 1 — наличие.

Как можно видеть, алгоритм нацелен на быструю фильтрацию шумовых примеров (не содержащий искомого шаблона). Фильтры Хаара формирующие каскад подбираются с помощью алгоритма AdaBoost имеющего похожую структуру, пороги $\tau^i$ обычно подбираются таким образом, что доля пропущенных положительных примеров в пределах наперед заданной доли.

Каскады Виолы-Джонса являются хорошим кандидатом для осуществления первичной фильтрации событий. Однако, как можно видеть каскады Виолы-Джонса не являются инвариантными ни к сдвигу изображения, ни к изменению масштаба. Учет сдвига изображения часто осуществляется с помощью применения многократного алгоритма к изображениям, сдвинутыми всевозможным образом с шагом $\Delta/2$, где $\Delta$ — характерный размер детектируемого шаблона; в тоже время предполагается, что сдвиги меньшие $\Delta/2$ учитываются самим алгоритмом. Изменение масштаба не является проблемой для эксперимента CRAYFIS, так как ожидаемая плотностью пикселей в камерах мобильных телефонов варьируется незначительно, что означает, что характерный размер следов космического излучения примерно одинаков для разных камер.

## 3.2 Ленивые вычисления в сверточных сетях

Однако, при решении проблемы первичной фильтрации мы сталкиваемся с еще одним препятствием — различными характеристиками камер телефонов. Одной из таких характеристик является глубина и структура пикселей камеры, которая значительно влияет на отклик детектора. Например, при увеличении глубины пикселей, ионизирующее излучение проходит через

большее количество материала, имея возможность внести больше энергии, т.е. породить больше электронов. Таким образом, частица с низкой энергией может оставить такой же след, как и частица с высокой энергией, но проходящая через более тонкие пиксели, несмотря на то, что следы могут быть легко отличимы при использовании одной и той де камеры. Понятно, что это может существенно деградировать качество восстановления параметров событий. Подход, позволяющий учесть различия в характеристиках камеры описан в последующих главах. Этот подход требует более гибкой модели триггера, а именно полную дифференцируемость модели триггера, что на практике оставляет немного опций. В этой работе мы рассматривает алгоритмы глубинного обучения, а именно сверточные нейронные сети.

Использование классических сверточных нейронных сетей в качестве алгоритма триггера является затруднительным, так как типичные вычислительные мощности не позволяют запуск моделей достаточно больших для достижения адекватных результатов. В этой работе мы предлагаем использования нового метода — комбинации каскадов и сверточных нейронных сетей.

Основная идея заключается в том, чтобы модифицировать таким образом, чтобы предоставить возможность остановки вычислений в шумовых регионах изображения. Это достигается за счет:

- введения карт активации;
- модификации алгоритма оператора свертки.

Сверточные нейронные сети в процессе работы вычисляют последовательность уменьшающихся изображений. Каждое такое изображение, обычно, содержит множество каналов (оригинальное изображение обычно содержит либо одним, черно-белый канал, либо три — каждый для кодирования красного, зеленого и синего цветов), соответствующим каким-либо высокоуровневым признакам (например, наличию в данном участке изображения линии, частей лица и проч.). Для каждого промежуточного изображения мы вводим дополнительной изображение такого же размера, но с одним каналом — карту активации. Значения карты активации соответствуют степени уверенности в наличии искомого шаблона в соответствующем участке изображения — если это значение падает ниже определенного порога, вычисления в данном участке прекращаются. Алгоритм оператора свертки модифицируется таким образом, чтобы учитывать это значение, не производя вычисления в регионе, если соответствующее значение карты активации ниже порога.

На карты активации накладываются дополнительные свойства, для того, чтобы соответствовать основной идее алгоритма, а именно возможности прервать вычисления — в одном и том же регионе, значения карты активации никогда не повышаются при переходе от одного слоя сверточной сети к другому. Следующие уравнения описывают вычисления карт активации:

$$A^0[x,y] = 1; \tag{3.10}$$

$$A^t[x,y] = \begin{cases} \sigma(W_A^t \cdot F^t[x,y]) & \text{если } A^{t-1}[x,y] > \tau; \\ A^{t-1}[x,y] & \text{иначе;} \end{cases} \tag{3.11}$$

где: $A^t$ — массив, представляющий карту активации после слоя $t$, $F^t[x,y] \in \mathbb{R}^{m^t}$ — промежуточные изображения, получаемые сверточной нейронной сетью, $\sigma$ — сигмоидальная функция активации.

Изначально все изображение "активировано что отражает предположение, что позиции событий равновероятны. Стоит заметить, что при нарушении этого предположения (например, если некий регион камеры является дефектным), изначальная карта активации может быть модифицирована, чтобы отразить априорное знание.

Структура одного ленивого сверточного оператор иллюстрирована на рисунке 3.1, вся структура алгоритма триггера на рисунке 3.2.

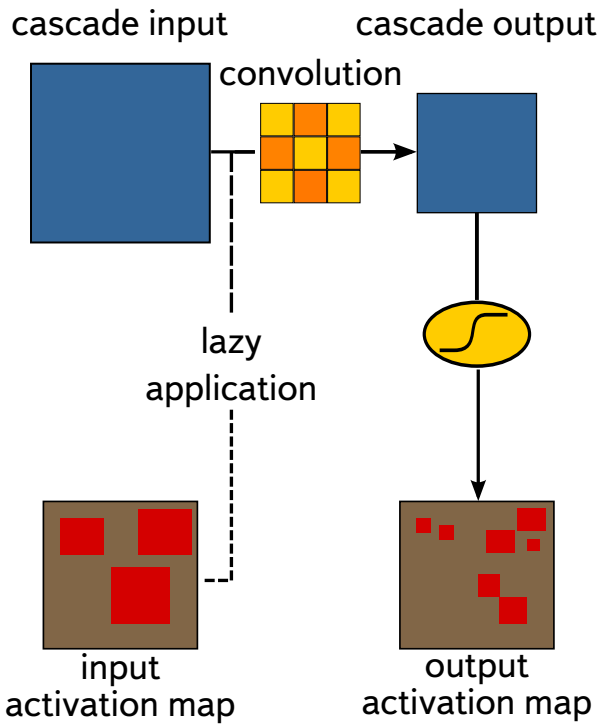



Рис. 3.1: Структура ленивого сверточного оператора. Левая часть изображения соответствует выходам предыдущего ленивого сверточного оператора.

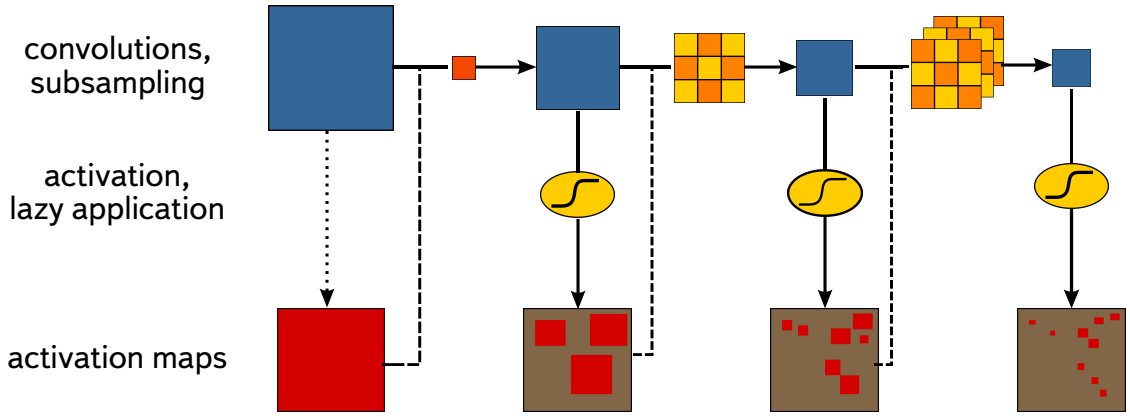



Рис. 3.2: Структура ленивой сверточной сети. Изначальная карта активации заполнена значениями 1. В процессе работы активированные регионы сужаются.

## 3.3 Обучение ленивой сверточной сети

Полное прекращение вычислений представляет собой проблему для обучения, так как для обучения сверточных нейронных сетей используются градиентные алгоритмы, но прекращение вычислений не является дифференцируемым действие. Однако, стоит заметить, что из-за не возрастания значений карт активации, возможно замена ленивой свертки на классическую (т.е. свертку без учета значений карты активации). В этом случае на время тренировки сети, вычисление карт активации производится по следующей схеме:

$$A^0[x,y] = 1; \tag{3.12}$$
$$A^t[x,y] = \sigma(W_A^t \cdot F^t[x,y]) \cdot A^{t-1}[x,y]; \tag{3.13}$$

что с одной стороны обеспечивает дифференцируемость, с другой сохраняет свойство невозрастания.

Тренировка сети происходит минимизацией среднего значения кросс-энтропии по всем регионам изображения:

$$L_2 = \sum_{i,j} y_{i,j} \log A[i,j] + (1 - y_{i,j}) \log(1 - A[i,j]); \tag{3.14}$$

где: $A[i,j]$ обозначает карту активации последнего слоя сверточной сети, $y_{i,j}$ — бинарный индикатор наличия искомого шаблона в регионе с координатами $(i,j)$.

Стоит заметить, что у функции потерь $L_2$ существует тривиальное решение, в котором все карты активации, кроме последней выдают значение 1, что фактически соответствует полному вычислению сети, что идет вразрез с основной целью алгоритма. Для того, чтобы найти вычислительно эффективное решение, предлагается использование следующего регуляризационного члена:

$$C = \sum_{k=1}^{n} c^k \sum_{i,j} (1 - y_{i,j})\, A_{i,j}^{k-1}; \tag{3.15}$$

где первая сумма проходит слои сверточной сети, $c^k$ — коэффициент соответствующий сложности вычисления значения фильтров слоя $k$. Фактически, регуляризация (3.15) является дифференцируемой аппроксимацией вычислительной сложности всей сети (на шумовых примерах) и позволяет напрямую влиять на сложность вычислений всей сети.

Итоговая функция потерь выглядит следующим образом:

$$L = \sum_{i,j} y_{i,j} \log A[i,j] + (1 - y_{i,j}) \log(1 - A[i,j]) + \lambda \left[ \sum_{k=1}^{n} c^k \sum_{i,j} (1 - y_{i,j})\, A_{i,j}^{k-1} \right]. \tag{3.16}$$

Рисунок 3.3 демонстрирует структуру ленивой сверточной сети во время обучения.

## 3.4 Эксперименты

Для того, чтобы продемонстрировать эффективность алгоритма, была сформирована искусственная обучающая выборка максимально приближенная к реальности. Для получения следов

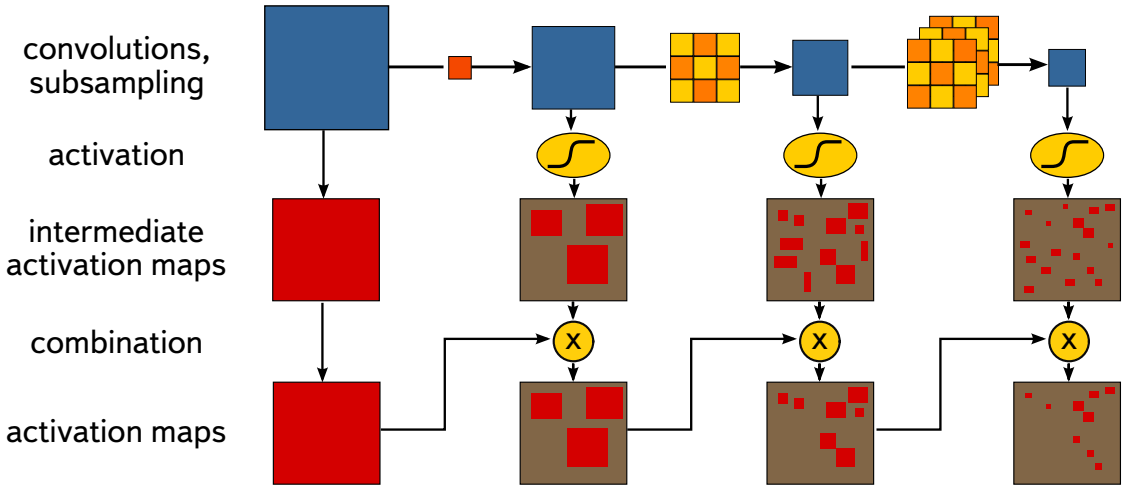
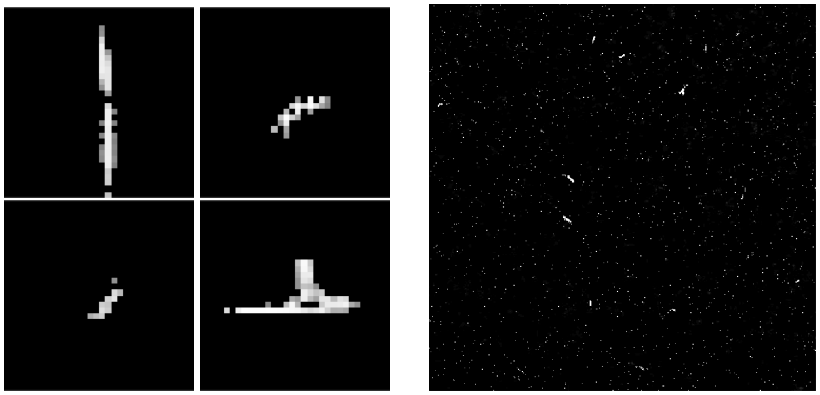



Рис. 3.3: Структура ленивой сверточной сети во время обучения.

частиц, телефон был подвергнут излучению радиоактивного источника $^{226}$Ra. Этот источник излучает фотоны рентгеновского спектра, которые при взаимодействии с материалом камеры порождают электроны. Последние в свою очередь оставляют детектируемые следы на камере мобильного телефона. Из физических рассуждений, эти следы должны быть эквивалентны следам от космического излучения с единственной разницей в яркости — электроны оставляют яркие, легко выделяемые следы. Поэтому для формирования обучающей выборки, следы электронов были выделены из экспериментальных изображений и их яркость было понижена до уровня соответствующему шуму камеры. Для большей реалистичности, полученные изображения были наложены на изображения шума камеры. Рисунок 3.4 иллюстрирует процесс создания обучающей выборки.



(a) оригинальные следы частиц (b) композиция

Рис. 3.4: Иллюстрация шагов формирования обучающей выборки: 3.4a выбор ярких следов, оставленных высоко-энергичными фотонами, 3.4b выбранные изображения собраны в композицию.

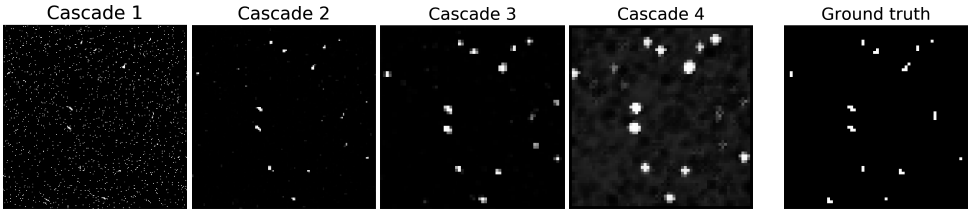
В эксперименте использовалась 4-х слойная ленивая сверточная сеть, обученная с функцией потерь (3.16). В таблице 3.1 приводятся результаты двух сетей, обученных с разными коэффициентами $\lambda$.

Качество полученное сетью со средней сложностью 2.0 операции на пиксель сравнимы с качеством достигаемым эквивалентной сверточной сетью без ленивых сверток, требуя при этом около 4-5 процентов вычислительной мощности.

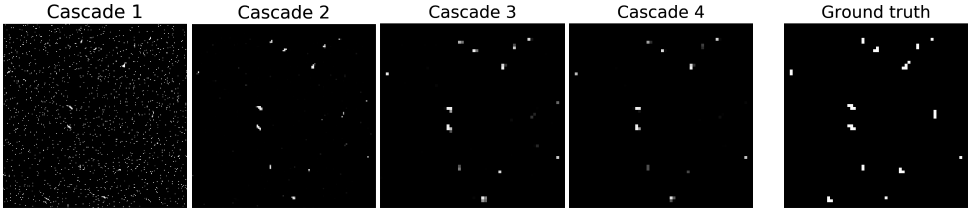
| complexity | 1.4 op. per pixel | | | 2.0 op. per pixel | | |
|---|---|---|---|---|---|---|
| signal efficiency | 0.90 | 0.95 | 0.99 | 0.90 | 0.95 | 0.99 |
| background rejection | 0.60 | 0.39 | 0.12 | 0.65 | 0.44 | 0.15 |

Таблица 3.1: Качество ленивых сверточных сетей со средними значениями вычислительной сложности 1.4 и 2.0 операции на пиксель.

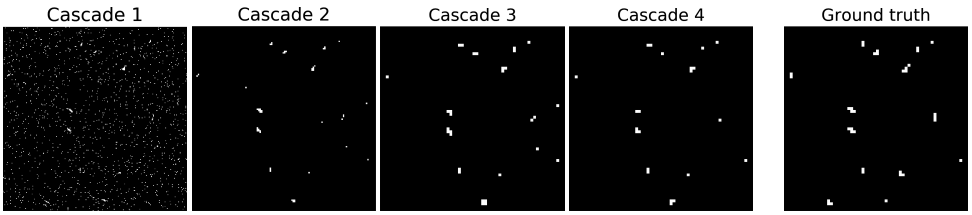
Рисунок 3.5 показывает карты активации полученные на разных слоях сети.




(a) промежуточные карты активации и настоящие значения




(b) карты активации и настоящие значения




(c) бинаризированные карты активации и настоящие значения

Рис. 3.5: Примеры карт активации.

## 3.5 Вывод

В данной главе были предложены ленивые сверточные сети, на основе которых строится триггер для эксперимента CRAYFIS. Триггер является неотъемлемой частью системы обработки данных и от его производительности зависит успех эксперимента. Было показано, что предложенная архитектура ленивых сверточных сетей достигает высокой скорости обработки данных, что позволяет запуск триггера на самих мобильных телефонов. Это существенно уменьшает нагрузку на сервера эксперимента и значительно уменьшает стоимость самого эксперимента.

Стоит отметить, что предложенная архитектура может быть применена и для других задач обнаружения редких событий на изображениях.

# Глава 4

# Обучение на реальных данных

Одной из часто встречающихся проблем в астрофизике и физике высоких энергий является невозможность получения меток для реальных данных для тренировки алгоритмов классификации. В данной главе мы ограничимся рассмотрением бинарной классификации, где один класс условно называется шумом, другой сигналом. В некоторых случаях допустимо использование компьютерных симуляций для получения данных и меток, однако зачастую результаты симуляции систематически отличаются от наблюдаемых данных, поэтому классификаторы натренированные на искусственных субоптимальны [26, 27].

Самым популярным подходом является использование техники sPlot [28] является тренировка на контрольной переменной. Контрольной переменной в данном случае называют наблюдаемую величину, которая:

- статистически независима от остальных наблюдаемых величин внутри каждого из классов;
- имеет известные распределения для каждого из классов.

Стоит заметить, что, в отличии от точных меток, контрольные переменные намного чаще встречаются на практике.

Техника sPlot [28] использует контрольную переменную для вычисления весов sWeights для каждого примера. Основное свойство sWeights состоит в том, что для любой функции $f(x)$, взвешенное среднее по всей тренировочной выборке, является несмещенной оценкой мат. ожидания $f(x)$ по сигнальной выборке:

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{sw}_i \cdot f(x_i) \approx \mathop{\mathbb{E}}_{x\sim S} f(x); \tag{4.1}$$

$$\frac{1}{N}\sum_{i=1}^{N} (1-\mathrm{sw}_i) f(x_i)) \approx \mathop{\mathbb{E}}_{x\sim B} f(x); \tag{4.2}$$

где: $S$, $B$ — обозначают распределения сигнального и шумового классов, $\mathrm{sw}_i$ — веса, вычисленные по технике sPlot.

Наивное применение к задача классификации ведет к следующей функции потерь:

$$L(\theta) = -\sum_{i=1}^{N} \mathrm{sw}_i \log f_\theta(x_i) + (1-\mathrm{sw}_i)\log\left(1-f_\theta(x_i)\right); \tag{4.3}$$

где: $f_\theta$ — классификатор с параметрами $\theta$. Несложно заметить, что мат. ожидания функций (4.3) и кросс-энтропии с точными метками классов равны.

В работе [26] было сделано важное наблюдение, в большинстве случаев в обучающей выборке присутствуют примеры с $\mathrm{sw}_i < 0$ или $\mathrm{sw}_i > 1$, что означает, что для достаточно мощного классификатора, способного изолировать данные примеры, функция (4.3) теряет нижнюю границу, что может привести к абсурдным результатам. Данный феномен иллюстрирует рисунок 4.1: стоит заметить, что обучение нейронной сети на функции потерь (4.3) нестабильно и ведет к субоптимальным результатам, в то время как идентичные сети обученные с помощью других функций потерь стабильны и не переобучены.

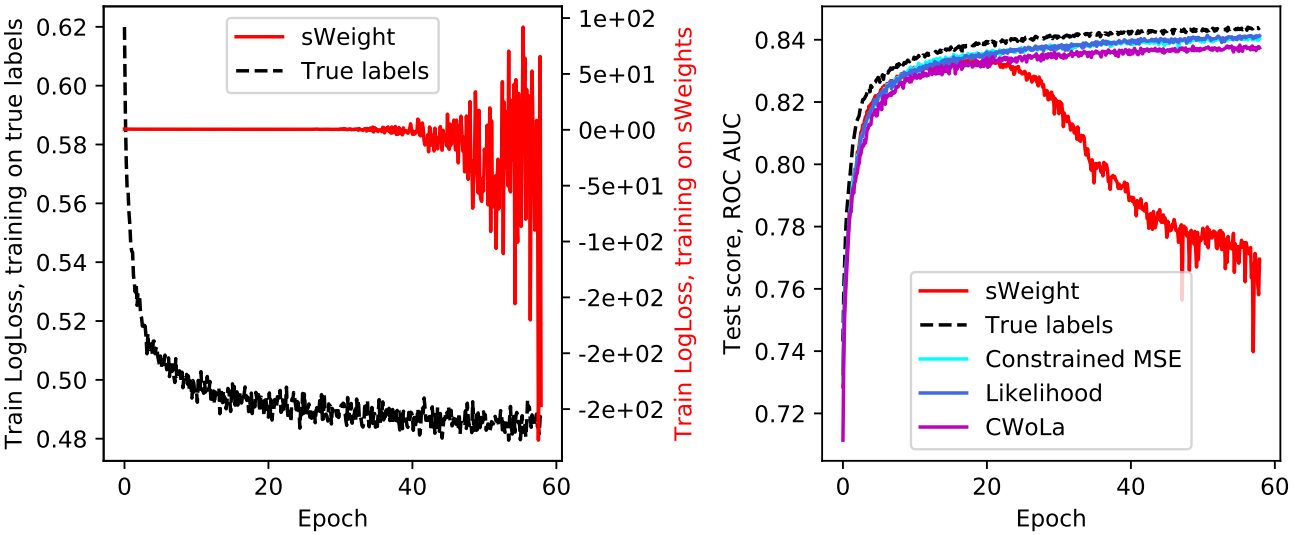



Рис. 4.1: композиция

Рис. 4.2: Кривые обучения для нейронных сетей на наборе данных HIGGS с искусственной контрольной переменной. Слева: значение функции потерь на тренировочной выборке в зависимости от эпохи обучения. Справа: ROC AUC на тестовой выборке в зависимости от эпохи обучения.

Для решения этой проблемы, были предложены две альтернативные функции потерь:

$$L(\theta) = \sum_i \left( \mathrm{sw}_i - \frac{e^{f_\theta(x_i)}}{1 + e^{f_\theta(x_i)}} \right)^2; \tag{4.4}$$

$$L(\theta) = -\sum_i \log \left[ f_\theta(x_i) p_\mathrm{s}(m_i) + (1 - f_\theta(x_i)) p_\mathrm{b}(m_i) \right]; \tag{4.5}$$

где: $m_i$ — значение контрольной переменной для i-го примера, $p_b$, $p_m$ — плотности вероятности контрольной переменной для шумового и сигнального классов.

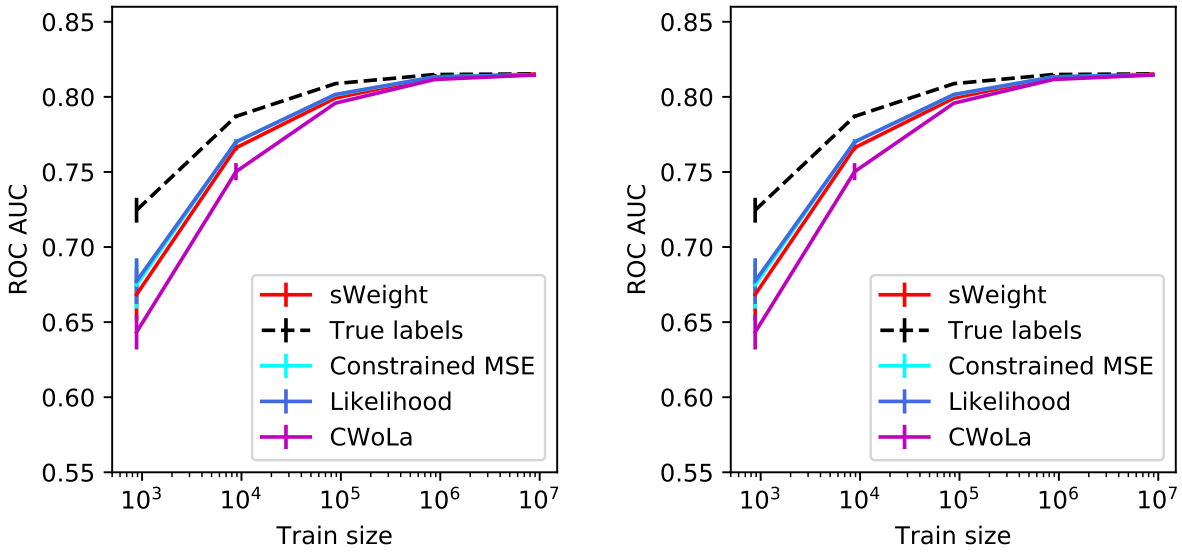
Первая функция потерь опирается на следующие утверждения:

- $E_{m_i} \mathrm{sw}_i = P(S \mid x_i)$;
- среднеквадратичные потери восстанавливают среднее значение метки в асимптотическом пределе;
- среднее значение sWeights по контрольной переменной заключено в отрезке $[0, 1]$.

Вторая функция потерь является правдоподобием и вместо sWeights использует оригинальные значения плотностей вероятности для контрольной переменной.

## 4.1 Эксперимент

Предложенные функции потерь были протестированы на наборе данных Higgs с искусственной контрольной переменной. Для сравнения, также был обучен метод CWoLa [27]. Результаты представлены на рисунке 4.3.




(a) Градиентный бустинг.

(b) Нейронный сети.

Рис. 4.3: Результаты сравнения функций потерь на наборе данных Higgs. Горизонтальные оси показывают размер обучающей выборки. Вертикальные оси соответствует метрике ROC AUC на тестовых данных.

## 4.2 Приложение к эксперименту CRAYFIS

Для эксперимента CRAYFIS проблема тренировки без достоверных меток возникает при тренировке триггера, описанного в предыдущей главе — следы ионизирующего излучения практически невозможно отделить от шума камеры с абсолютной достоверностью, более того, ручная разметка изображений (даже при гипотетическом наличии некого достоверного критерия) была бы крайне трудоемка.

В этом случае в качестве контрольной переменной используется отношение яркости самого яркого пикселя к яркости самого яркого его соседа — из физических соображений, значения шума в двух пикселях статистически независимы, в то время как следы частиц непрерывны и значения сильно скоррелированы. Распределения отношения яркости для шума и сигнала можно достоверно получить из реальных данных (для шума) и симуляции (для сигнала).

Однако, такая контрольная переменная не является независимой от изображения. Данная проблема решается с помощью удаления информации о контрольной переменной из промежуточного представления нейронной сети с помощью метода [23]. Таким образом, несмотря на зависимость контрольной переменной от изображения, нейронная сеть принимает решение по представлению изображения, которое уже не зависит от контрольной переменной.

## 4.3 Вывод

Триггер, описанный в предыдущей главе, является неотъемлемой частью системы обработки данных для эксперимента CRAYFIS. Предложенные в этой главе методы позволяют стабильную тренировку триггера на реальных наблюдаемых данных без привлечения дополнительных экспериментов с целью получения достоверных меток классов, что существенно снижает затраты на проведения эксперимента.

Стоит отметить, что метод не является специфичным для эксперимента CRAYFIS, и может быть применен для любых задач классификации с контрольной переменной. Такие задачи крайне часто встречаются на практике в астрофизике и физике высоких энергий. Также стоит отметить, что предложенные функции потерь достигают наилучших результатов по сравнению с методами тренировки на контрольной переменной, описанными в литературе.

# Глава 5

# Автоматическая верификация качества данных

Конечной цель эксперимент CRAYFIS является обнаружение ливней космического излучения ультра-высоких энергий (Ultra-High Energy Cosmic Rays, UHECR), которые, как уже было замечено во введении, являются крайне редкими событиями по сравнению с частотой событий космического фона. С точки зрения обработки данных, основная особенность обнаружения ливней UHECR заключается в том, что их поведение заранее неизвестно, поэтому UHECR следует рассматривать как аномалии по отношению к типичному космическому фону.

Обнаружение аномалий также важно при оценке качества данных поступающих от волонтеров, так как точная проверка соблюдения всех условий эксперимента не представляется возможным. Например, неизвестно для волонтера, телефон может находиться рядом с локальным источником радиации, что заметно повысит частоту событий, детектируемых телефоном, по сравнению с космическим фоном. В качестве другого примера можно привести проникновение света, вызванное плохо закрытой камерой.

С учетом вышесказанного, основной целью эксперимента является нахождение аномалий (ливней UHECR) среди аномальных данных (по сравнению с космическим фоном), что делает методы обнаружения аномалий в данных одной из важнейших задач для эксперимента CRAYFIS. Стоит также отметить, что формальная постановка в данном случае расходится с классическими определениями детектированием аномалий (Anomaly Detection) и обнаружением выбросов (Outlier Detection) в анализе данных. Главной причиной тому, является ценность аномалий, и сложность формального доказательства правильной работы любого алгоритма обучения без учителя. Поэтому в данной работе под обнаружением аномалий понимается задача классификации (т.е. обучение с учителем). Обучение классификатора происходит на метках предоставленных экспертами.

Предлагаемая система обнаружения аномалий для эксперимента CRAYFIS была спроектирована по подобию системы верификации качества данных эксперимента LHC CMS. Эксперименты, показывающие эффективность предлагаемых алгоритмов так же проверялись на данных полученных экспериментом LHC CMS.

Особенность экспериментов CRAYFIS и LHC CMS является огромный (предполагаемый) поток данных. В эксперименте LHC CMS, для целей верификации качества, данные группируются примерно по 23 секунды работы детектора. Для каждый группы данных вычисляются многочисленные статистики, большинство из которых мотивированных предметной областью.

Для оценки качества одной группы данных от эксперта требуется проанализировать около сотни гистограмм, и, в случае обнаружения несоответствий, определить возможные причины из дополнительных источников. В случае эксперимента CRAYFIS ожидается похожий процесс

оценки качества данных, но в этом случае, объем данных ожидается на порядки выше из-за большого ожидаемого количества независимых детекторов.

Указанные выше причины являются мотивацией к созданию автоматической системы оценки качества данных, которая может обучаться на разметки экспертов, постепенно принимая больше и больше автоматических решений.

В работе [36] вводиться алгоритм автоматической верификации данных. Основная идея алгоритма состоит в том, чтобы обучать алгоритм классификации по меткам предоставленным экспертами, и принимать автоматические решение только в случае, если алгоритм выдает достаточно высокую вероятность события быть либо аномальным, либо нормальным.

Более формально, классификатор $f$ разделяет данные $x$ на три категории:

- пример автоматически классифицируется как нормальный: $f(x) > \tau_L$;
- решение передается эксперту: $\tau_P \leqslant f(x) \leqslant \tau_L$;
- пример автоматически классифицируется как аномальный: $f(x) < \tau_P$;

где $\tau_L$, $\tau_R$ — пороги автоматического решения. При передачи решения эксперту, после получения метки, размеченный пример добавляется в обучающую выборку и классификатор обучается заново.

Пороги автоматического решения, технически, могут быть фиксированы заранее. Однако, в общем случае, ограничения на следующие величины представляют собой практический интерес:

$$\text{Pollution Rate} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Positive}} \leqslant P_0; \tag{5.1}$$

$$\text{Loss Rate} = \frac{\text{False Negative}}{\text{False Negative} + \text{True Positive}} \leqslant L_0;; \tag{5.2}$$

где $P_0$ и $L_0$ — ограничения заданные заранее. Первая величина отвечает за уровень «загрязнения» данных, количество аномальных данных среди примеров размеченных как положительные, вторая величина определяет потери нормальных данных. Стоит отметить, что при расчете величин (5.1) и (5.2) учитываются только автоматически размеченные данные.

При обучении классификатора, пороги $\tau_L$ и $\tau_P$ подбираются таким образом, чтобы решения классификаторы удовлетворяли ограничениям на Loss; Rate и Pollution Rate. Так как эти величины определяются только по автоматически принятым решениям, для любых пределов $P_0$ и $L_0$ и любого классификатора существуют константы $\tau_L$ и $\tau_P$, такие что ограничения удовлетворены. Конечно, при $\tau_P = 0$ и $\tau_L = 1$ ограничения (5.1) и (5.2) будут всегда удовлетворены, однако такие пороги снизят количество автоматических решений до нуля, что противоречит основной цели системы — минимизации работы эксперта:

$$\text{Rejection Rate} = \frac{\text{Rejected}}{\text{Total}} \to \min; \tag{5.3}$$

где: Rejected — количество примеров переданных эксперту, Total — общее количество примеров, переданных в системе. В связи с этим пороги $\tau_P$ и $\tau_L$ подбираются как решения задачи (5.3) при ограничениях (5.1) и (5.2). При этом Loss; Rate и Pollution Rate оцениваются с помощью кросс-валидации. Полная процедура автоматической верификации качества данных представлена в алгоритме 1.

**Algorithm 1** Алгоритм автоматической оценки качества данных

**function** Train($X$, $y$, $L_0$, $P_0$)
  compute scores $\hat{y}$ by $k$-fold cross-validation
  $\tau_L = \max\{\tau \mid \hat{L}_\tau(\hat{y}, y) \leqslant L_0\}$
  $\tau_P = \min\{\tau \mid \hat{P}_\tau(\hat{y}, y) \leqslant P_0\}$
  **return** $\tau_L$, $\tau_P$, classifier trained on $X, y$
**end function**

**function** AutomatedDataQuality($L_0$, $P_0$)
  $\tau_L, \tau_P \leftarrow 0, 1$
  classifier $\leftarrow \frac{1}{2}$
  $X_{\text{train}} = \varnothing$
  $y_{\text{train}} = \varnothing$
  **for** $i = 0, 1, \ldots, N$ **do**
    $x_i \leftarrow$ new sample
    $\hat{y}_i \leftarrow$ classifier$(x_i)$
    **if** $\hat{y}_i > \tau_L$ **then**
      classify $x_i$ as good lumisection
    **else if** $\hat{y}_i < \tau_P$ **then**
      classify $x_i$ as anomalous lumisection
    **else**
      $y_i \leftarrow$ label from human expert
      $X \leftarrow (X, x_i)$
      $y \leftarrow (y, y_i)$
      $\tau_L, \tau_P$, classifier $\leftarrow$ Train$(X, y, L_0, P_0)$
    **end if**
  **end for**
**end function**

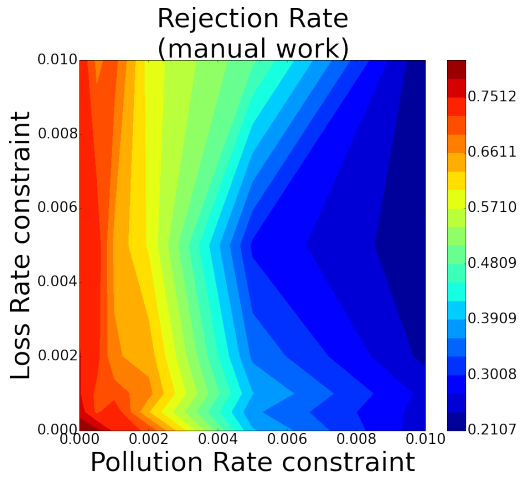
## 5.1 Эксперимент

Для оценки качества работы предлагаемого алгоритма верификации качества данных использовались данные собранные экспериментом LHC CMS. Данные разделены на так называемые люмисекции, каждая люмисекция содержит события, наблюдаемые в течении около 23 секунд. Для каждой люмисекции, были вычислены статистики мотивированные предметной областью, таким образом каждая люмисекция представляется в виде вектора фиксированной длины. Стоит отметить, что используемые статистики являются расширенным набором статистик, по которым эксперты принимают решения.

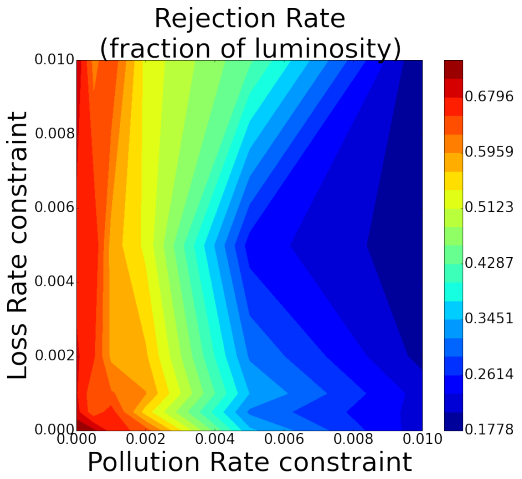
В качестве классификатора использовался Gradient Boosting над решающими деревьями.

На рисунке 5.1 приведены результаты работы алгоритма, измеренные в доле примеров, переданных эксперту, и в доле светимости событий, переданных эксперту (светимость в данном случае является эмпирическим показателем «важности» люмисекции, соответственно используется как вес). Как видно из результатов, алгоритм способен сохранять от 20% до 80% времени эксперта, при разумных ограничениях на Loss; Rate и Pollution Rate. Стоит отметить, что даже при самых строгих ограничениях ($L_0 = P_0 = 0$), алгоритм все еще способен автоматически принимать около 20% решений.

Рисунок 5.2 демонстрирует другую важную особенность алгоритма — качество работы алгоритма увеличивается со временем из-за увеличения обучающей выборки. На основании этого




(a) Rejection Rate в количествах люмисекций.




(b) Rejection Rate как доля светимости.

Рис. 5.1: 5.1a — доля люмисекций переданных эксперту, 5.1b — доля светимости переданная эксперту.

результата, можно утверждать, по пришествию некоторого времени, практически все решения будут приниматься автоматически, конечно, за исключением примеров, точное определение аномальности которых невозможно по предоставленным данным.

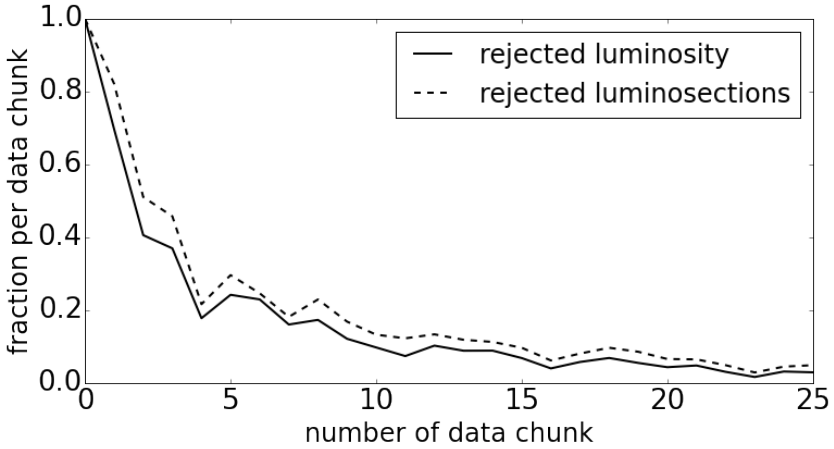



Рис. 5.2: Пример кривой обучения, количество люмисекций (и светимости этих люмисекций) переданных эксперту в зависимости от времени работы алгоритма.

## 5.2 Вывод

Предложенный в данной главе алгоритм является важной частью системы обработки данных для эксперимента CRAYFIS, так как позволяет убедиться в качестве данных, с минимальными трудозатратами со стороны экспертов.

Стоит отметить, что разработанные методы не являются специфичными для эксперимента CRAYFIS и могут применяться для решения широкого спектра задач, в частности, данный метод был успешно протестирован на открытых данных эксперимента CERN CMS.

# Глава 6

# Детектирование аномалий

После верификации и агрегации данных, начинается шаг обнаружения событий, кандидатов на ливни космического излучения сверх-высоких энергий. Особенность этого шага заключается в том, что получение репрезентативной выборки ливней затруднено, во-первых из-за вычислительной сложности симуляции, во-вторых из-за неизвестной природы частиц сверх-высоких энергий и соответственно ливней, ими порожденных (напомним, что главная цель эксперимента — наблюдение ранее упомянутых ливней). Из физических соображений выборку событий, вызванных космическим фоном, можно считать репрезентативной и многочисленной (ее относительно легко симулировать из-за независимости устройств при фоновых событиях).

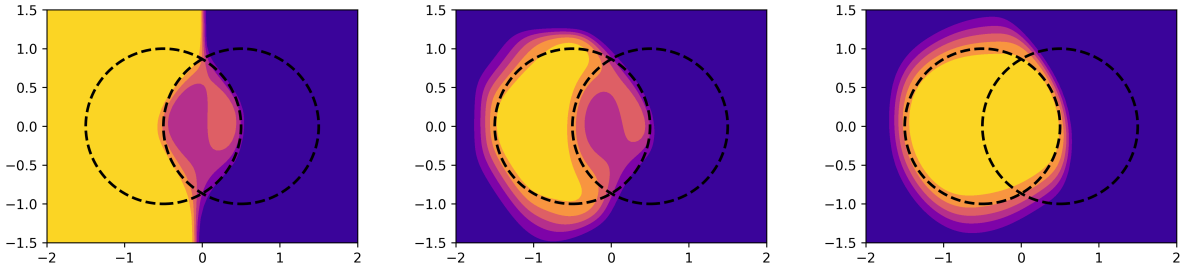
Данная постановка крайне похожа по описанию на классическую задачу обнаружения аномалий, в этом контексте, искомые ливни-кандидаты является аномалиями по отношению к космическому фону. Однако, эта постановка представляет собой особый класс задач:

- метки классов достоверно известны;
- выборка крайне не сбалансирована по классам из-за редкости ливней космического излучения сверх-высоких энергий;
- выборку аномалий нельзя считать полной: некоторые типы ливней могут не присутствовать в выборке из-за неизвестной природы;
- нельзя предполагать, что классы идеально отделимы.

В связи с вышеперечисленными особенностями, ни классические методы обнаружения аномалий, ни классические методы классификации не являются достаточными для решения задачи. Несложно видеть, что оптимальный метод должен обладать следующими свойствами:

- принимать решение близкое к классификации в областях с достаточным количеством данных;
- предсказывать аномальный класс в областях с отсутствующими данными;

другими словами, оптимальный метод является гибридом между двух-классовой и одной-классовой классификациями. Рисунок 6.1 иллюстрирует вышесказанное на примере синтетических данных.



(a) двух-классовая классификация (b) гибридный метод классификации (c) одно-классовая классификация

Рис. 6.1: Синтетический пример классификации. Классы равномерно распределены внутри окружностей обозначенных пунктирными линиями. Левый класс объявлен нормальным.

В данной работе были разработано семейство методов обладающих такими свойствами. Основная идея состоит в модификации классической функции потерь — кросс-энтропии:

$$\begin{aligned} \mathcal{L}_{1+\varepsilon}(f) &= \frac{1}{2}\left(L^+(f) + \gamma\, L^-(f) + (1-\varepsilon)\, L^0(f)\right); \\ L^+(f) &= -\mathop{\mathbb{E}}_{x\sim\mathcal{C}^+} \log f(x); \\ L^-(f) &= -\mathop{\mathbb{E}}_{x\sim\mathcal{C}^-} \log(1-f(x)); \\ L^0(f) &= -\mathop{\mathbb{E}}_{x\sim U} \log(1-f(x)); \end{aligned} \tag{6.1}$$

где $\mathcal{C}^+, \mathcal{C}^-$ — нормальный и аномальные классы, $f$ — функция классификатора, $U$ — равномерное распределение покрывающее $\operatorname{supp}\mathcal{C}^+ \cup \operatorname{supp}\mathcal{C}^-$. В работе [29] доказывается, что функция потерь (6.1) ведет к решению, обладающему вышеупомянутыми свойствами.

Для высокоразмерных данных, оценка функция потерь (6.1) может быть затруднительной из-за наличия равномерного сэмплирования по всему пространству. В работе [29] также вводиться альтернативная функция потерь:

$$\begin{aligned} \mathcal{L}^E_{1+\varepsilon}(f) &= \frac{1}{2}\left(L^+(f) + \gamma L^-(f) + (1-\varepsilon) L^E(\sigma^{-1}(f))\right); \\ \text{где: } L^E(g) &= \log Z = \int_\Omega \exp(g(x))dx; \\ \frac{1}{1+\exp(-g(x))} &= f(x); \end{aligned} \tag{6.2}$$

Основное отличие функции потерь (6.2) от (6.1) стоит в том, что теперь градиент регуляризационного члена вычисляется по точкам гораздо более узкого распределения $P_g$:

$$\begin{aligned} \nabla L^E(g) &= \nabla \log Z = \frac{1}{Z}\int_\Omega \exp(g(x))\nabla g(x) = \mathop{\mathbb{E}}_{x\sim P_g} \nabla g(x); \\ \text{где: } P_g(x) &= \frac{1}{Z}\frac{f(x)}{1-f(x)} = \frac{1}{Z}\exp(g(x)); \\ Z &= \int_\Omega \exp(g(x))dx. \end{aligned} \tag{6.3}$$

Желаемые свойства решения $\mathcal{L}^E_{1+\varepsilon}$ доказываются в работе [29].

| | one class | 100 | 1000 | 10000 | 1000000 |
|---|---|---|---|---|---|
| Robust AE | $0.530 \pm 0.002$ | $0.530 \pm 0.002$ | $0.530 \pm 0.002$ | $0.530 \pm 0.002$ | $0.530 \pm 0.002$ |
| Deep SVDD | $0.497 \pm 0.006$ | $0.497 \pm 0.006$ | $0.497 \pm 0.006$ | $0.497 \pm 0.006$ | $0.497 \pm 0.006$ |
| cross-entropy | - | $0.496 \pm 0.017$ | $0.529 \pm 0.007$ | $0.566 \pm 0.006$ | $0.858 \pm 0.002$ |
| semi-supervised | - | $0.498 \pm 0.003$ | $0.522 \pm 0.003$ | $0.603 \pm 0.002$ | $0.745 \pm 0.005$ |
| brute-force OPE | $0.499 \pm 0.009$ | $0.500 \pm 0.009$ | $0.520 \pm 0.003$ | $0.572 \pm 0.005$ | $0.859 \pm 0.001$ |
| HMC EOPE | $0.491 \pm 0.000$ | $0.523 \pm 0.005$ | $\mathbf{0.567} \pm 0.008$ | $\mathbf{0.648} \pm 0.005$ | $0.848 \pm 0.001$ |
| RMSProp EOPE | $0.498 \pm 0.002$ | $0.494 \pm 0.008$ | $0.531 \pm 0.008$ | $0.593 \pm 0.011$ | $\mathbf{0.861} \pm 0.000$ |
| Deep EOPE | $\mathbf{0.531} \pm 0.000$ | $\mathbf{0.537} \pm 0.011$ | $0.560 \pm 0.008$ | $0.628 \pm 0.005$ | $0.860 \pm 0.001$ |

Рис. 6.2: Результаты на данных HIGGS, первая строка обозначает количество примеров аномального класса использованных для тренировки.

| | one class | 100 | 1000 | 10000 | 1000000 |
|---|---|---|---|---|---|
| Robust AE | $0.394 \pm 0.012$ | $0.394 \pm 0.012$ | $0.394 \pm 0.012$ | $0.394 \pm 0.012$ | $0.394 \pm 0.012$ |
| Deep SVDD | $0.541 \pm 0.022$ | $0.541 \pm 0.022$ | $0.541 \pm 0.022$ | $0.541 \pm 0.022$ | $0.541 \pm 0.022$ |
| cross-entropy | - | $0.658 \pm 0.033$ | $0.736 \pm 0.021$ | $0.757 \pm 0.036$ | $0.871 \pm 0.006$ |
| semi-supervised | - | $0.715 \pm 0.020$ | $0.766 \pm 0.009$ | $\mathbf{0.847} \pm 0.002$ | $0.876 \pm 0.000$ |
| brute-force OPE | $\mathbf{0.648} \pm 0.035$ | $0.678 \pm 0.025$ | $0.729 \pm 0.029$ | $0.757 \pm 0.036$ | $0.871 \pm 0.006$ |
| HMC EOPE | $0.472 \pm 0.000$ | $\mathbf{0.738} \pm 0.019$ | $\mathbf{0.770} \pm 0.012$ | $0.816 \pm 0.006$ | $0.877 \pm 0.000$ |
| RMSProp EOPE | $0.443 \pm 0.038$ | $0.714 \pm 0.019$ | $0.760 \pm 0.016$ | $0.807 \pm 0.004$ | $0.877 \pm 0.000$ |
| Deep EOPE | $0.468 \pm 0.118$ | $0.670 \pm 0.054$ | $0.746 \pm 0.024$ | $0.813 \pm 0.003$ | $\mathbf{0.878} \pm 0.000$ |

Рис. 6.3: Результаты на данных SUSY, первая строка обозначает количество примеров аномального класса использованных для тренировки.

## 6.1 Эксперимент

Для оценки качества предложенного метода, было проведено ряд экспериментов. Для создания необходимых условий, из классических сбалансированных наборы данных были созданы подвыборки с различными соотношения классов. В случаях много-классовых наборов данных, также была взята подвыборка по классам (иными словами в наборе данных для тренировки присутствовали не все классы аномалий). Наборы данных участвовавших в эксперименте также включают классические наборы данных из физики высоких энергий (HIGGS, SUSY). В таблицах 6.2, 6.3 и 6.4 представлены результаты сравнения предложенных методов (brute-force OPE, HMC EOPE, RMSProp EOPE, Deep EOPE) с классической двух-классовой классификацией (cross-entropy), частичное обучение (сжатие пространства и двух-классовая классификация) и современных одно-классовых методов Robust AE [30] и Deep SVDD [31].

| | one class | 1 | 2 | 4 |
|---|---|---|---|---|
| Robust AE | $\mathbf{0.585} \pm 0.126$ | $0.585 \pm 0.126$ | $0.585 \pm 0.126$ | $0.585 \pm 0.126$ |
| Deep SVDD | $0.546 \pm 0.058$ | $0.546 \pm 0.058$ | $0.546 \pm 0.058$ | $0.546 \pm 0.058$ |
| cross-entropy | - | $0.659 \pm 0.093$ | $0.708 \pm 0.086$ | $0.748 \pm 0.082$ |
| semi-supervised | - | $0.587 \pm 0.109$ | $0.634 \pm 0.109$ | $0.671 \pm 0.093$ |
| brute-force OPE | $0.549 \pm 0.098$ | $\mathbf{0.688} \pm 0.087$ | $\mathbf{0.719} \pm 0.079$ | $\mathbf{0.757} \pm 0.073$ |
| HMC EOPE | $0.547 \pm 0.116$ | $0.678 \pm 0.091$ | $0.709 \pm 0.084$ | $0.739 \pm 0.074$ |
| RMSProp EOPE | $0.565 \pm 0.111$ | $0.678 \pm 0.081$ | $0.715 \pm 0.083$ | $0.746 \pm 0.069$ |
| Deep EOPE | $0.564 \pm 0.094$ | $0.674 \pm 0.100$ | $0.690 \pm 0.092$ | $0.719 \pm 0.099$ |

Рис. 6.4: Результаты на данных CIFAR-10, первая строка обозначает количество аномальных подклассов (из 9), входящих в обучающую выборку, из каждого аномального подкласса случайным образом выбраны по 10 примеров.

## 6.2 Вывод

В данной главе предлагается семейство методов позволяющих решить задачу обнаружения ливней-кандидатов для эксперимента CRAYFIS. Данный шаг является важной частью системы обработки данных, так как позволяет существенно снизить вычислительные ресурсы требуемые для точного анализа описанного в следующей главе.

Стоит отметить, что разработанные методы не являются специфичными для эксперимента CRAYFIS и могут применяться для решения широкого спектра задач.

# Глава 7

# Определение параметров камер мобильных телефонов

После отбора кандидатов в ливни, для точного подтверждения наблюдения космической части сверх-высокой энергии, требует точная реконструкция события. Так как ожидаемые количество событий и кандидатов низкое, то для точной реконструкции возможно использование вычислительных тяжелых методов. Основной проблемой при реконструкции события является неизвестные параметры камер телефонов. Для восстановления параметров камеры мобильных телефонов, участвовавших в событии-кандидате, предлагается нахождение соответствующих параметров симуляции, наиболее точно соответствующих историческим данным с мобильного телефона. Однако, симуляция и исторические данные стохастичны, формально данная задача описывается как нахождение параметров распределения, наилучшим образом описывающего наблюдаемые данные. Одним из наиболее успешных подходов является Adversarial Optimization, которая минимизирует дивергенцию Дженсона-Шеннона между распределением симулятора и распределением, заданным наблюдаемыми данными (с альтернативными подходами можно ознакомиться в ссылках в [32]). Для оценки дивергенции Дженсона-Шеннона используется классификатор:

$$\mathrm{JSD}_{\mathcal{F}}(P,Q) = \log 2 + \frac{1}{2}\max_{f\in\mathcal{F}}\left[\mathop{\mathbb{E}}_{x\sim P}\log f(x) + \mathop{\mathbb{E}}_{x\sim Q}\log(1-f(x))\right]; \tag{7.1}$$

где: JSD — дивергенция Дженсона-Шеннона, $\mathcal{F}$ — множество всех возможных классификаторов. Стоит заметить, что максимум в уравнение (7.1) берется по всем возможным классификаторам. На практике, множество $\mathcal{F}$ заменяется достаточно мощной моделью классификатора (например, нейронной сетью).

Несмотря на редкость событий-кандидатов, тренировка мощного классификатора (которая должна производиться на каждом шаге оптимизации) требует большого количества примеров симулятора, последний является вычислительно трудозатратным.

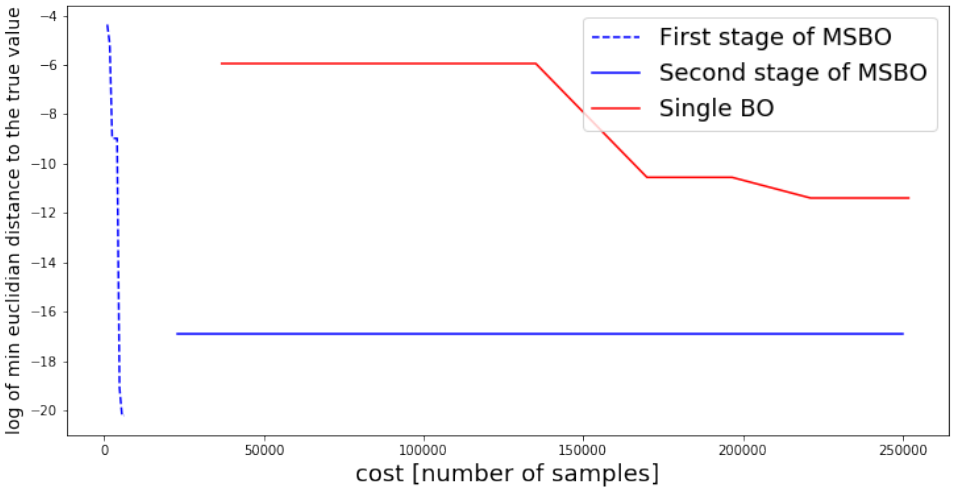
Заметим, что при замене множества $\mathcal{F}$ в уравнении (7.1) на простую модель $\mathcal{G}$, величина $\mathrm{JSD}_{\mathcal{G}}(P,Q)$ становится псевдо-дивергенцией, с единственным отличием от оригинальной дивергенции в том, что из $\mathrm{JSD}_{\mathcal{G}}(P,Q) = 0$ не следует равенство $P$ и $Q$. Однако, простые модели, как правило, требуют значительно меньшего количества примеров для тренировки, при этом все еще предоставляя информацию о нахождении решения для оптимизации.

Для того, чтобы воспользоваться данным свойством была предложена многоступенчатая Байесовская оптимизация [32]. Основная идея заключается в том, что на первом шаге алгоритм использует $\mathrm{JSD}_{\mathcal{G}}(P,Q)$, которая из-за простоты классификатора позволяет быстро просканировать все пространство параметров. Затем, вторая Байесовская оптимизация запускает-

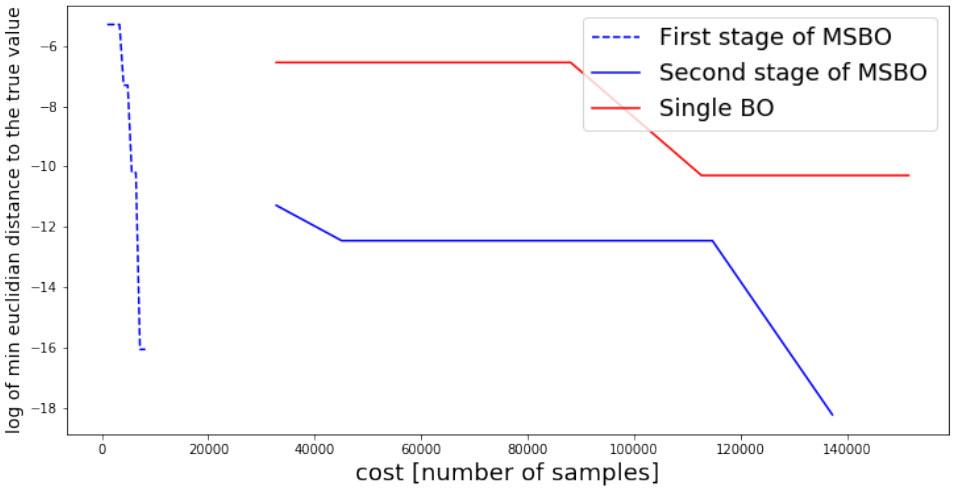
ся для нахождения минимума дивергенции Дженсона-Шеннона, оцененной достаточно мощным классификатором, но с ограничением на параметры $\mathrm{JSD}_{\mathcal{G}}(P, Q) = 0$. Из-за погрешности в оценки (псевдо-) дивергенций и вероятностной природы суррогатных моделей используемых в Баейсовской оптимизации, последние ограничения превращаются в следующее условие: $P(\mathrm{JSD}_{\mathcal{G}}(P, Q) < \varepsilon) > P_0$.

## 7.1 Эксперимент

Эффективность данного метода была оценена на задаче восстановления параметров генератора событий столкновения электрона-антиэлектрона Pythia. Данная задача была предложена в работе [33].




(a) 1 оптимизируемый параметр.




(b) 4 оптимизируемых параметра.

Рис. 7.1: Сравнение скорости сходимости Байесовская оптимизации (красная линия) с многоступенчатой Байесовской оптимизацией (первая ступень показаны пунктирной синей линией, вторая ступень показана непрерывной синей линией). Горизонтальная ось соответствует количеству запусков симуляции (суммарному размеру обучающих выборок). Вертикальная ось соответствует логарифму расстояния промежуточного решения оптимизатора до настоящих параметров.

Пример сравнения сходимости классической Байесовской оптимизации с оценкой дивергенции Дженсона-Шеннона с помощью достаточно мощного классификатора и предложенного алгоритма приведен на рисунке 7.1. Как можно видеть из рисунка, первая ступень оптимизации быстро сужает область поиска оптимума (в первом случае подходя к минимум ближе, чем вторая ступень), вторая ступень используется для уточнения положения минимума.

## 7.2 Вывод

В данной главе предложен алгоритм, применяемый в последнем шаге системы обработки данных в эксперименте CRAYFIS. Эксперимент показывает, что предложенный алгоритм позволяет существенно сэкономить вычислительные ресурсы при определении параметров симуляции по наблюдаемой выборке событий, что снижает стоимость всего эксперимента.

Как и остальные методы рассмотренные в данной работе, многоступенчатая Байесовская оптимизация может применяться для широкого спектра задача, выходящего за рамки конкретного эксперимента.

# Глава 8

# Заключение

В данной работе предложены основные алгоритмы для обработки данных для космической обсерватории на основе мобильных телефонов для наблюдения за космическим излучение сверх-высоких энергий (эксперимент CRAYFIS), а именно были рассмотрены следующие методы (соответствующие публикации и выступления на конференциях процитированы):

- построение быстрого триггера на основе ленивых сверточных сетей [34];
- обучение при отсутствии достоверных меток (с помощью контрольной переменной) [35];
- верификация качества данных [36];
- детектирования аномалий при наличии нерепрезентативной выборки известных аномалий [29].
- быстрое определение параметров генератора [32].

Эксперимент CRAYFIS в данной работе используется как полигон для демонстрации работоспособности предложенных алгоритмов. Сами алгоритмы имеют гораздо более широкий спектр применения в решении задач обработки данных в астрофизике и физике высоких энергий.

# Литература


[1] Correlation of the highest-energy cosmic rays with the positions of nearby active galactic nuclei / J Abraham, P Abreu, M Aglietta et al. // Astroparticle Physics. — 2008. — Vol. 29, no. 3. — P. 188–204.

[2] Bell A. The acceleration of cosmic rays in shock fronts–i // Monthly Notices of the Royal Astronomical Society. — 1978. — Vol. 182, no. 2. — P. 147–156.

[3] Waxman E. Cosmological origin for cosmic rays above 1019 ev // The Astrophysical Journal Letters. — 1995. — Vol. 452, no. 1. — P. L1.

[4] Weiler T. J. Cosmic-ray neutrino annihilation on relic neutrinos revisited: a mechanism for generating air showers above the greisen-zatsepin-kuzmin cutoff // Astroparticle Physics. — 1999. — Vol. 11, no. 3. — P. 303–316.

[5] Beane S. R., Davoudi Z., Savage M. J. Constraints on the universe as a numerical simulation // The European Physical Journal A. — 2014. — Vol. 50, no. 9. — P. 1–9.

[6] Cogliati J. J., Derr K. W., Wharton J. Using cmos sensors in a cellphone for gamma detection and classification // arXiv preprint arXiv:1401.0766. — 2014.

[7] Low cost, pervasive detection of radiation threats / Gordon A Drukier, Eric P Rubenstein, Peter R Solomon et al. // Technologies for Homeland Security (HST), 2011 IEEE International Conference on / IEEE. — 2011. — P. 365–371.

[8] Searching for ultra-high energy cosmic rays with smartphones / Daniel Whiteson, Michael Mulhearn, Chase Shimmin et al. // Astroparticle Physics. — 2016. — Vol. 79. — P. 1–9.

[9] James F. Statistical methods in experimental physics. — World Scientific, 2006.

[10] D'Agostini G. Bayesian reasoning in data analysis: A critical introduction. — World Scientific, 2003.

[11] Ball N. M., Brunner R. J. Data mining and machine learning in astronomy // International Journal of Modern Physics D. — 2010. — Vol. 19, no. 07. — P. 1049–1106.

[12] The lhcb detector at the lhc / A Augusto Alves Jr, LM Andrade Filho, AF Barbosa et al. // Journal of instrumentation. — 2008. — Vol. 3, no. 08. — P. S08005.

[13] LHCb Trigger and Online Upgrade Technical Design Report : Rep. : CERN-LHCC-2014-016. LHCB-TDR-016 / CERN. — Geneva : 2014. — May. — Access mode: https://cds.cern.ch/record/1701361.

[14] Whiteson S., Whiteson D. Machine learning for event selection in high energy physics // Engineering Applications of Artificial Intelligence. — 2009. — Vol. 22, no. 8. — P. 1203–1217.

[15] Lhcb topological trigger reoptimization : Rep. ; Executor: Philip Iten, Tatiana Likhomanenko, Egor Khairullin et al. : 2015.

[16] Gligorov V. V., Williams M. Efficient, reliable and fast high-level triggering using a bonsai boosted decision tree // Journal of Instrumentation. — 2013. — Vol. 8, no. 02. — P. P02013.

[17] Learning to discover: the higgs boson machine learning challenge / Claire Adam-Bourdarios, Glen Cowan, Cecile Germain et al. // URL http://higgsml. lal. in2p3. fr/documentation. — 2014.

[18] Sadowski P. J., Whiteson D., Baldi P. Searching for higgs boson decay modes with deep learning // Advances in Neural Information Processing Systems. — 2014. — P. 2393–2401.

[19] Egmont-Petersen M., de Ridder D., Handels H. Image processing with neural networks—a review // Pattern recognition. — 2002. — Vol. 35, no. 10. — P. 2279–2301.

[20] Schmidhuber J. Deep learning in neural networks: An overview // Neural Networks. — 2015. — Vol. 61. — P. 85–117.

[21] Baldi P., Sadowski P., Whiteson D. Searching for exotic particles in high-energy physics with deep learning // Nature communications. — 2014. — Vol. 5.

[22] New approaches for boosting to uniformity / Alex Rogozhnikov, Aleksandar Bukva, Vladimir Gligorov et al. // Journal of Instrumentation. — 2015. — Vol. 10, no. 03. — P. T03002.

[23] Ganin Y., Lempitsky V. Unsupervised domain adaptation by backpropagation // Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37 / JMLR. org. — 2015. — P. 1180–1189.

[24] Girshick R. Fast r-cnn // Proceedings of the IEEE international conference on computer vision. — 2015. — P. 1440–1448.

[25] Faster r-cnn: Towards real-time object detection with region proposal networks / Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun // Advances in neural information processing systems. — 2015. — P. 91–99.

[26] Borisyak M., Kazeev N. Machine learning on data with sPlot background subtraction // Journal of Instrumentation. — 2019. — aug. — Vol. 14, no. 08. — P. P08020–P08020.

[27] Metodiev E. M., Nachman B., Thaler J. Classification without labels: Learning from mixed samples in high energy physics // Journal of High Energy Physics. — 2017. — Vol. 2017, no. 10. — P. 174.

[28] Pivk M., Le Diberder F. R. Plots: A statistical tool to unfold data distributions // Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. — 2005. — Vol. 555, no. 1-2. — P. 356–369.

[29] $(1 + \varepsilon)$-class classification: an anomaly detection method for highly imbalanced or incomplete data sets / Maxim Borisyak, Artem Ryzhikov, Andrey Ustyuzhanin et al. // arXiv preprint arXiv:1906.06096. — 2019.

[30] Zhou C., Paffenroth R. C. Anomaly detection with robust deep autoencoders // Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining / ACM. — 2017. — P. 665–674.

[31] Deep one-class classification / Lukas Ruff, Nico Görnitz, Lucas Deecke et al. // International Conference on Machine Learning. — 2018. — P. 4390–4399.

[32] Adversarial event generator tuning with bayesian optimization / Maxim Borisyak, Radoslav Neychev, Denis Derkach, Andrey Ustyuzhanin. — conference talk at Computing in High Energy Physics 2018, 2018.

[33] Ilten P., Williams M., Yang Y. Event generator tuning using bayesian optimization // Journal of Instrumentation. — 2017. — Vol. 12, no. 04. — P. P04028.

[34] Muon trigger for mobile phones / Maxim Borisyak, Michail Usvyatsov, Michael Mulhearn et al. // Journal of Physics: Conference Series / IOP Publishing. — Vol. 898. — 2017. — P. 032048.

[35] Borisyak M., Kazeev N. Machine learning on data with sPlot background subtraction // Journal of Instrumentation. — 2019. — aug. — Vol. 14, no. 08. — P. P08020–P08020.

[36] Towards automation of data quality system for cern cms experiment / Maxim Borisyak, Fedor Ratnikov, Denis Derkach, Andrey Ustyuzhanin // Journal of Physics: Conference Series / IOP Publishing. — Vol. 898. — 2017. — P. 092041.